Programming Project 2

Due date: Fri, Dec. 13, 2013 at 4pm.

Task description: The ability to search literature and collect/aggregate metrics around publications is a central tool for modern research. Both academic and industry researchers across hundreds of scientific disciplines, from astronomy to zoology, increasingly rely on search to understand what has been published and by whom.

Microsoft Academic Search is an open platform that provides a variety of metrics and experiences for the research community, in addition to literature search. It covers more than 50 million publications and over 19 million authors across a variety of domains, with updates added each week. One of the main challenges of providing this service is caused by author-name ambiguity. This task challenges participants to determine which authors in a given data set are duplicates. **Note:** This competition is identical to the Author Disambiguation Challenge that was hosted by KDD 2013.

Submission file: For every author in the dataset, submission files should contain two columns: *AuthorId* and *DuplicateAuthorIds*. The column *DuplicateAuthorIds* should be a space-delimited list. Every *AuthorId* counts as his/her own duplicate, and every duplicate should be listed under each of its respective ids. For example, if you suspect author `A`, `B`, and `C` are the same, you should list (`A`,`A B C`), (`B`,`B A C`), (`C`,`C A B`). The file should contain a header and have the following format:

```
AuthorId,DuplicateAuthorIds
1,1
8,8
9,9 10
10,10 9
etc.
```

Dataset description: The dataset(s) for the challenge are provided by Microsoft Corporation and come from their Microsoft Academic Search (MAS) database. MAS is a free academic search engine that was developed by Microsoft Research, and covers more than 50 million publications and over 19 million authors across a variety of domains.

- Author: is a publication author in the Academic Search dataset.
- Paper: is a scholarly contribution written by one or more authors - could be of type conference or journal. Each paper also has additional metadata, such as year of publication, venue, keywords, etc.

- Affiliation: the name of an organization with which an author can be affiliated.

The provided datasets are based on a snapshot taken in Jan 2013 and contain:

- An Author dataset (**Author.csv**) with profile information about 250K authors, such as author name and affiliation. The same author can appear more than once in this dataset, for instance because he/she publishes under different versions of his/her name, such as J. Doe, Jane Doe, and J. A. Doe.

Name	Data Type	Comments
Id	Int	Id of the author
Name	Nvarchar	Author name
Affiliation	Nvarchar	Organization name with which the author is affiliated

- A Paper dataset (**Paper.csv**) with data about 2.5M papers, such as paper title, conference/journal information, and keywords. The same paper may have been obtained through different data sources and hence have multiple copies in the dataset.

Name	Data Type	Comments
Id	Int	Id of the paper
Title	Nvarchar	Title of the paper
Year	Int	Year of the paper
ConferenceId	Int	Conference Id in which paper was published
JournalId	Int	Journal Id in which paper was published
Keywords	Nvarchar	Keywords of the paper

- A corresponding Paper-Author dataset (**PaperAuthor.csv**) with (paper ID, author ID) pairs. The Paper-Author dataset is noisy, containing possibly incorrect paper-author assignments that are due to author name ambiguity and variations of author names.

Name	Data Type	Comments
PaperId	Int	Paper Id
AuthorId	Int	Author Id
Name	Nvarchar	Author Name (as written on paper)

- Since each paper is either a conference or a journal, additional metadata about conferences and journals is provided where available (Conference.csv, Journal.csv).

Name	Data Type	Comments
Id	Int	Conference Id or Journal Id
ShortName	Nvarchar	Short name
FullName	Nvarchar	Full name
Homepage	Nvarchar	Homepage URL of conf/journal

Comments:

1. Co-authorship can be derived from the Paper-Author dataset
2. Files **Train.csv** and **Valid.csv** were part of the other track of the KDD Cup 2013, so you may ignore them

Evaluation: The goal of this competition is to predict which authors are duplicates. The task is structured as a "cold start" problem, meaning there are no training labels provided. Participants must develop their own duplicate criteria.

The evaluation metric for this competition is Mean `F1`-Score. The `F1` score, commonly used in information retrieval, measures accuracy using the statistics precision p and recall r. Precision is the ratio of true positives (tp) to all predicted positives (tp + fp). Recall is the ratio of true positives to all actual positives (tp + fn). The `F1` score is given by:

$$F1 = 2\frac{p \cdot r}{p + r} \quad \text{where} \quad p = \frac{tp}{tp + fp}, \quad r = \frac{tp}{tp + fn}$$

The `F1` metric weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favored over extremely good performance on one and poor performance on the other.

Since the majority of authors are not duplicates, please note that the `F1` score will be close to `1` for this task. To assess competition progress, you may wish to compare your `F1` score to the benchmark representing the "null" prediction (each author is his own duplicate). Small differences in the absolute magnitude of the `F1` score can represent meaningful improvements in model performance, despite our natural inclination to assume lesser decimal places are insignificant.

Writeup: In addition to submitting your solution online, you need to provide us with a 2-page writeup that describes the algorithm you have implemented and the special tricks you used in order to make it work. Also, describe your strategy for selecting that particular algorithm and how you did your offline evaluation of the method.

Instructions: If you don't have an account for Kaggle, you will need to signup. Go to `http://inclass.kaggle.com` and use your BU email. After that, search for *BU CS 565 : Project 2* and you will find the competitions.

If you encounter any problems, send an email at *cmav@bu.edu*